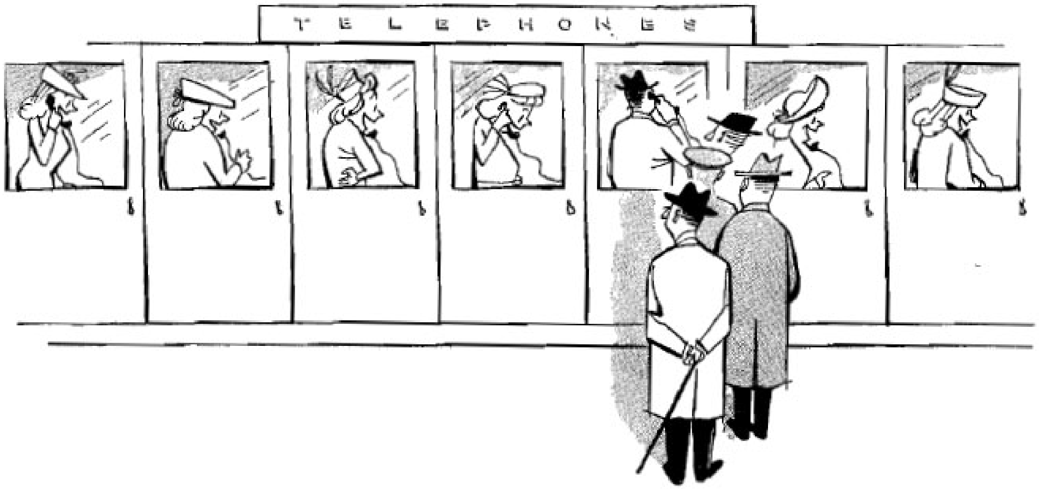
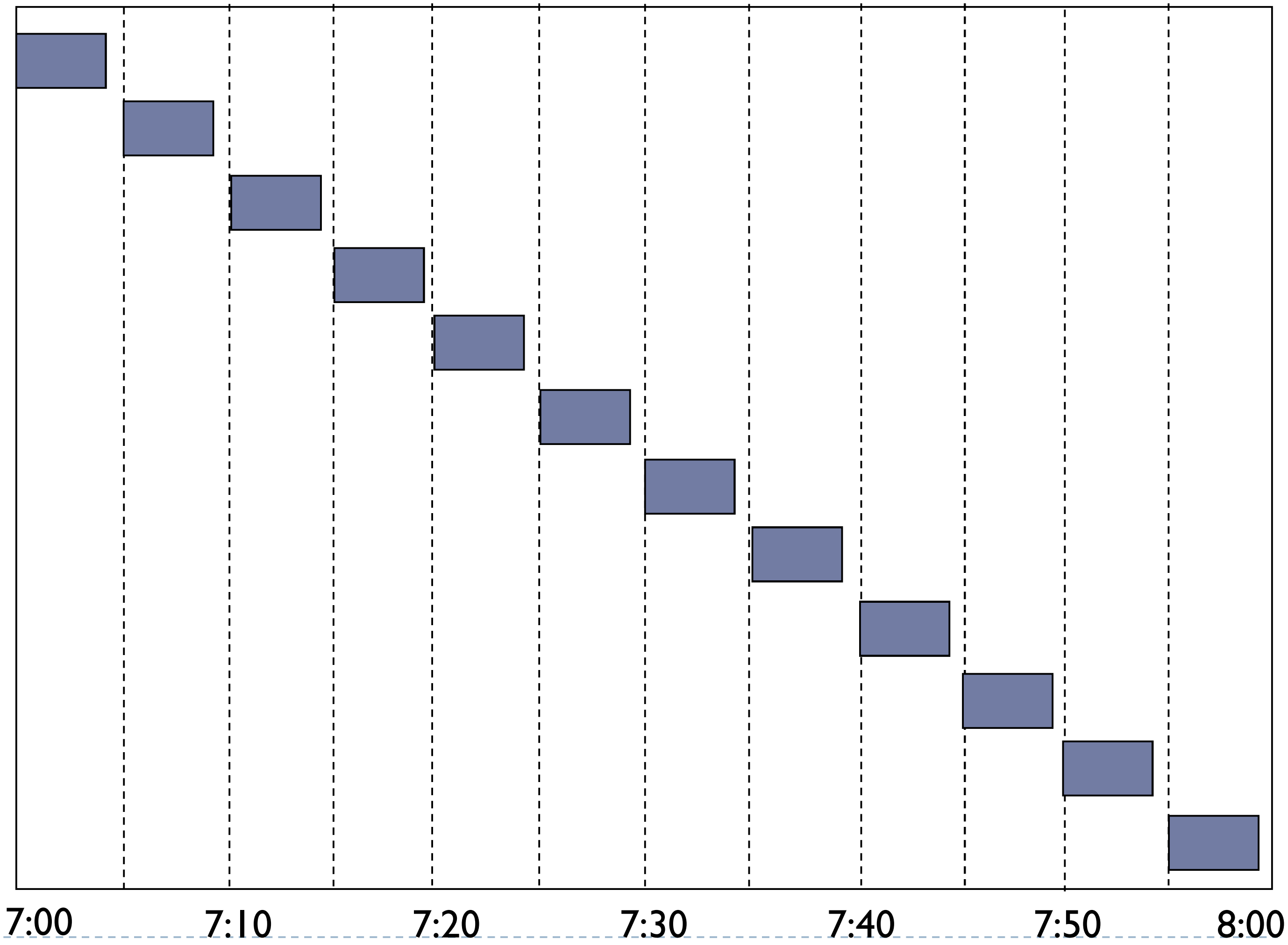


Modelos de Colas y Tiempos de Espera

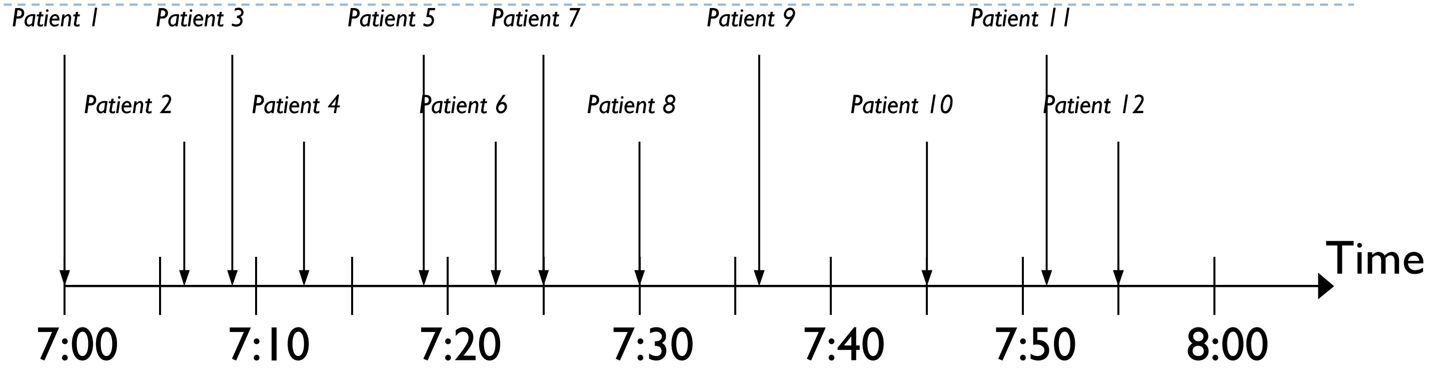
A Somewhat Odd Service Process

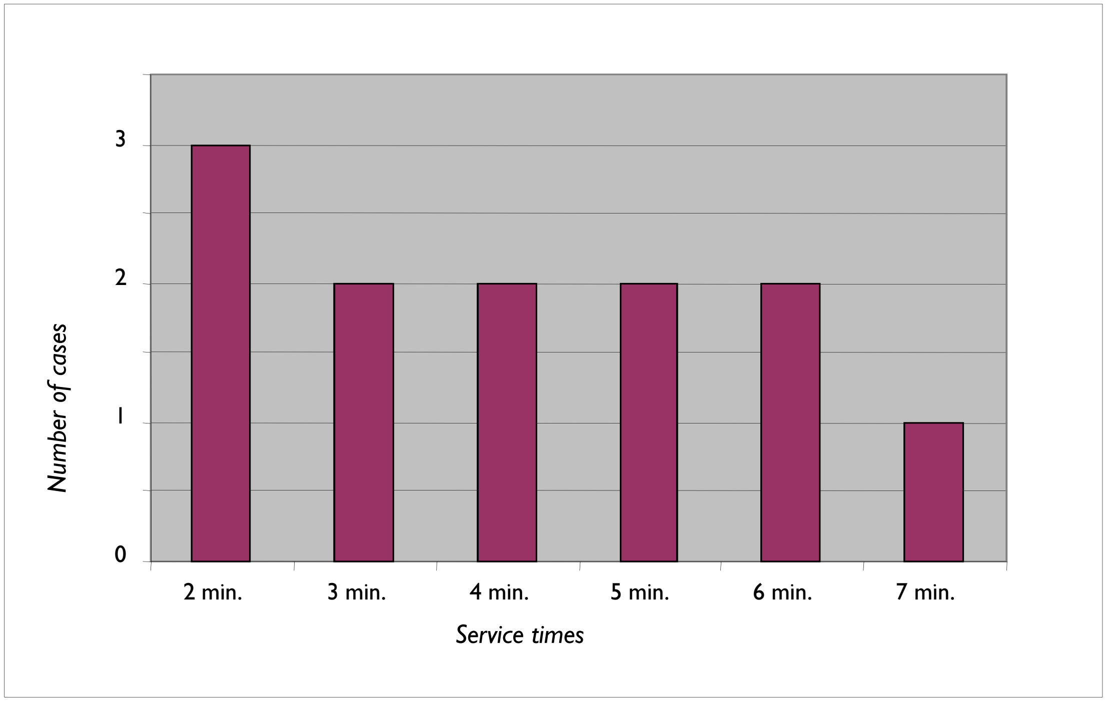
Patient	Arrival Time	Service Time
1	0	4
2	5	4
3	10	4
4	15	4
5	20	4
6	25	4
7	30	4
8	35	4
9	40	4
10	45	4
11	50	4
12	55	4



A More Realistic Service Process

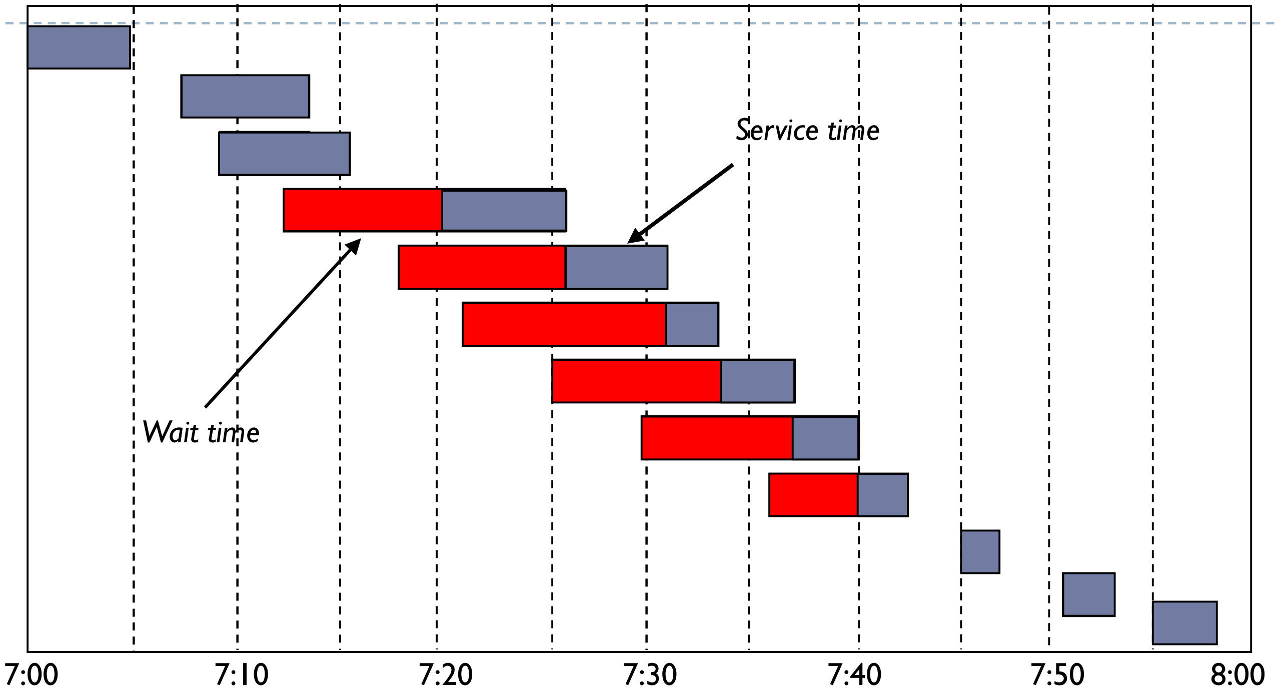
Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3

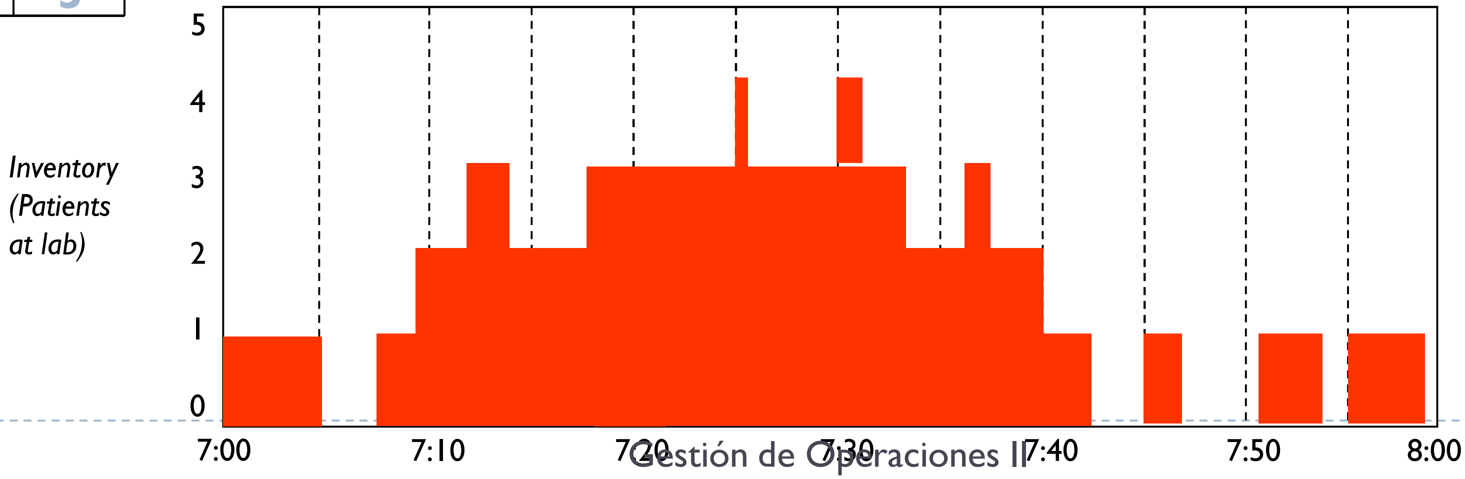




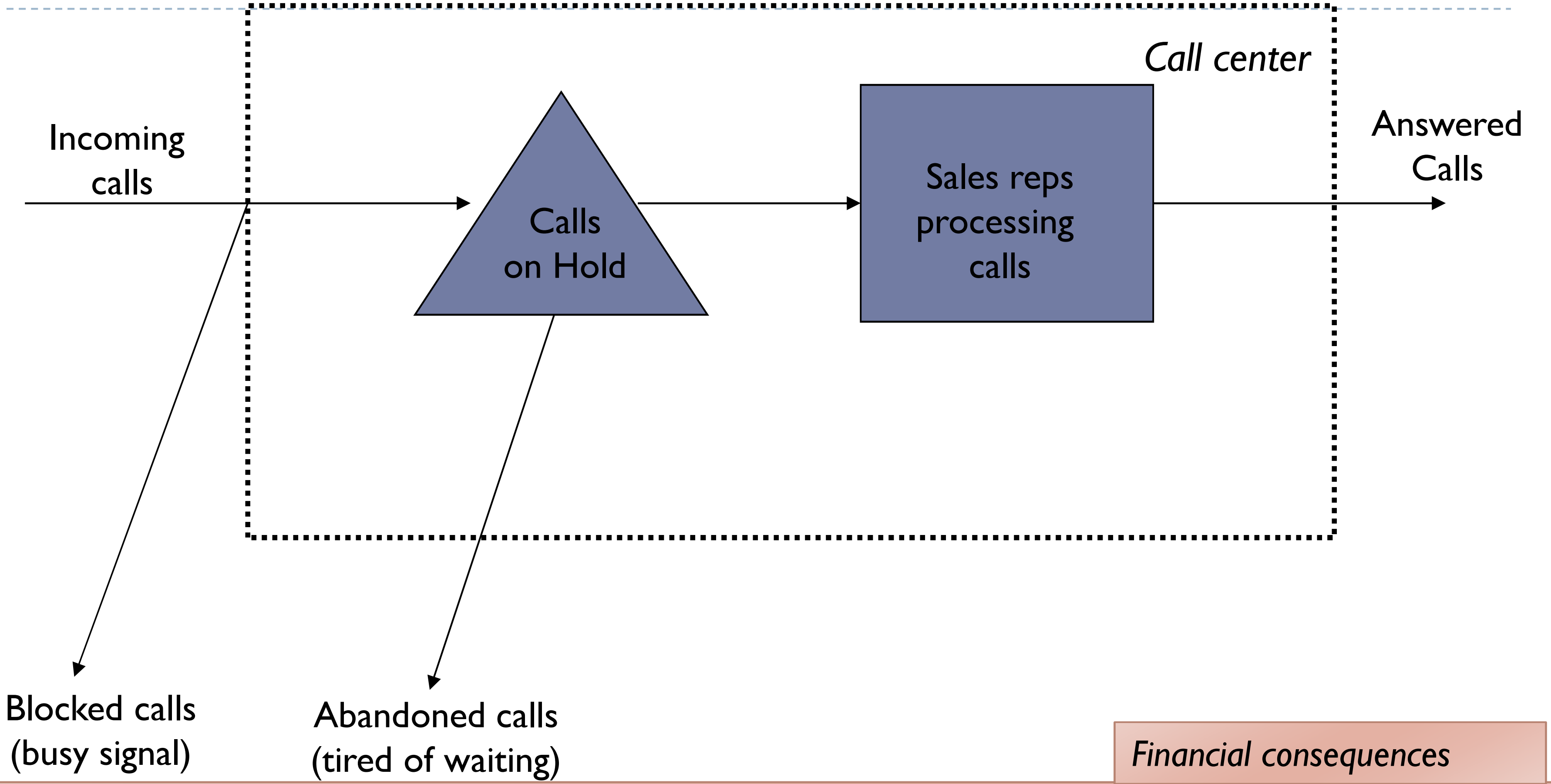
Variability Leads to Waiting Time… and Inventory

Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3





Example of a Queuing System: Call Centers



Call center

Incoming calls

Calls on Hold

Sales reps processing calls

Answered Calls

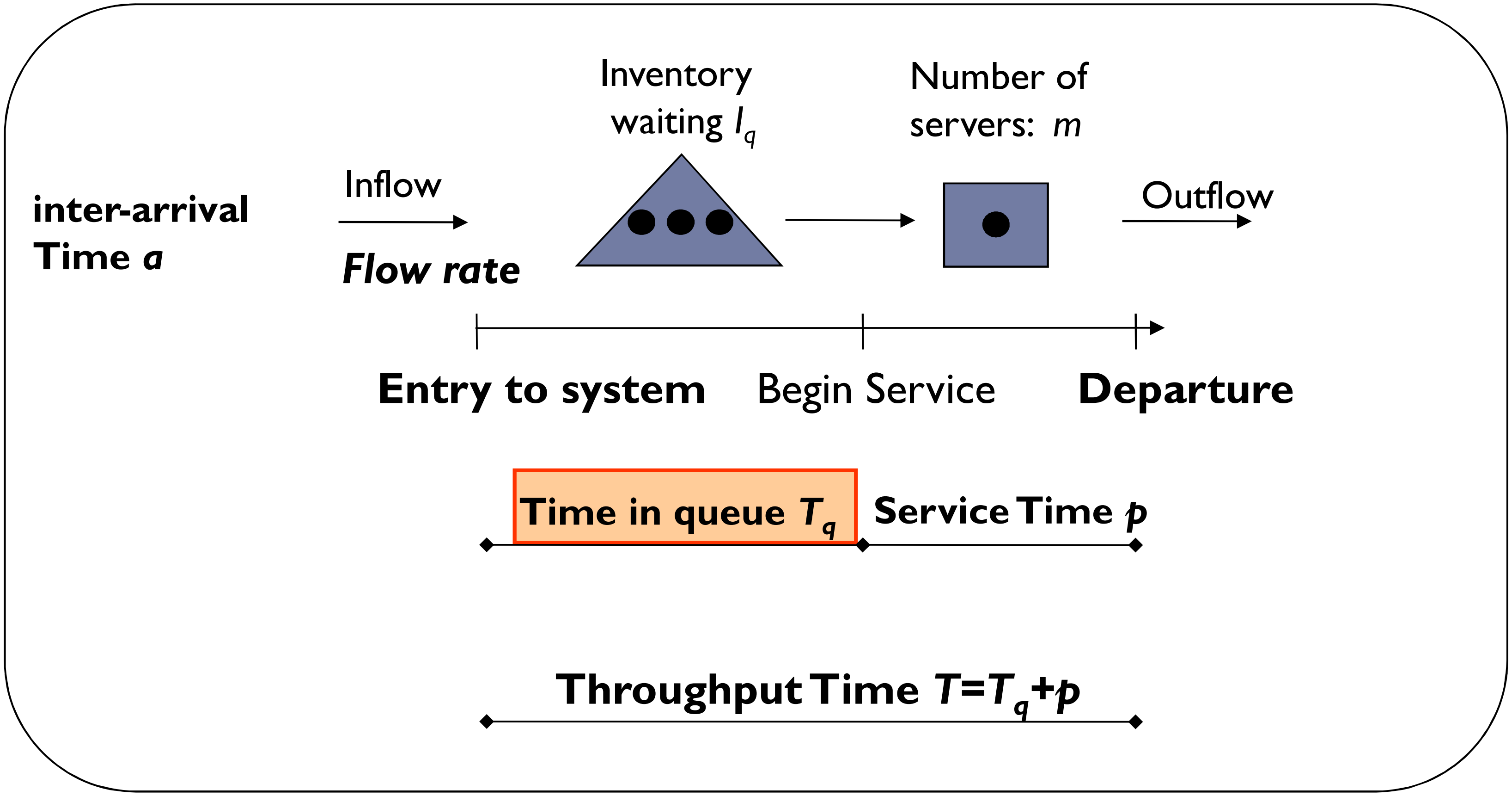
Blocked calls (busy signal)

Abandoned calls (tired of waiting)

Financial consequences

Lost throughput	Holding cost Lost goodwill Lost throughput (abandoned)	Cost of capacity Cost per customer	$$$ Revenue $$$

Queueing system model



Service Level Metrics

- Many service organizations define a service standard as an *Acceptable Waiting Time* which serves as an upper bound for the waiting time experienced by a given (large) percentage of customers, the *Service Level*:
- AWT= *Acceptable Waiting Time*= maximum amount of waiting time (in queue) experienced by SL% of the customers
- SL = *Service Level*=percentage of customers whose waiting time is at or below the AWT

Example; many call centers are designed such that SL= 80% or SL=90% of all customers are served within an AWT of 20 seconds

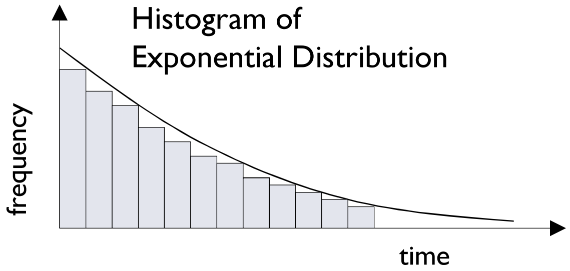
Most contractual agreements with outsourced call centers specify a *Service Level Agreement (SLA)* of this type

- T_q = *Mean* Waiting Time in *Queue*
- T_s = *Mean* Waiting Time in *System*
- N_q = *Mean* Number of Customers in *Queue*= demand rate* T_q (Little's Law)
- N_s = *Mean* Number of Customers in *System*= demand rate* T_s (Little's Law)
- P_d= probability of delay= likelihood a customer experiences any waiting time

Modeling Arrival and Service times

- To incorporate variability, an accurate queuing model generally requires a detailed description of the statistical distribution of arrivals and service times.

  Example:
  Time between consecutive arrivals follow an *Exponential* distribution (a.k.a."Poisson" arrivals)



- *Coefficient of Variation (CV):* measures the variability of a random variable X.

$$CV_X = \left(\frac{\text{standard deviation (X)}}{\text{mean (X)}}\right) \begin{array}{l} \nearrow \; CV_a : \text{arrival times} \\ \searrow \; CV_p : \text{service times} \end{array}$$

Example:
If the time between consecutive arrivals follow an Exponential distribution, then CVa=1 (this is a special property of the Exponential distribution).

The Erlang Model *

Basic assumptions:

(a) A pool of service agents with identical skills and characteristics

(b) Customers are serviced on a FIFO basis ; no priority classes

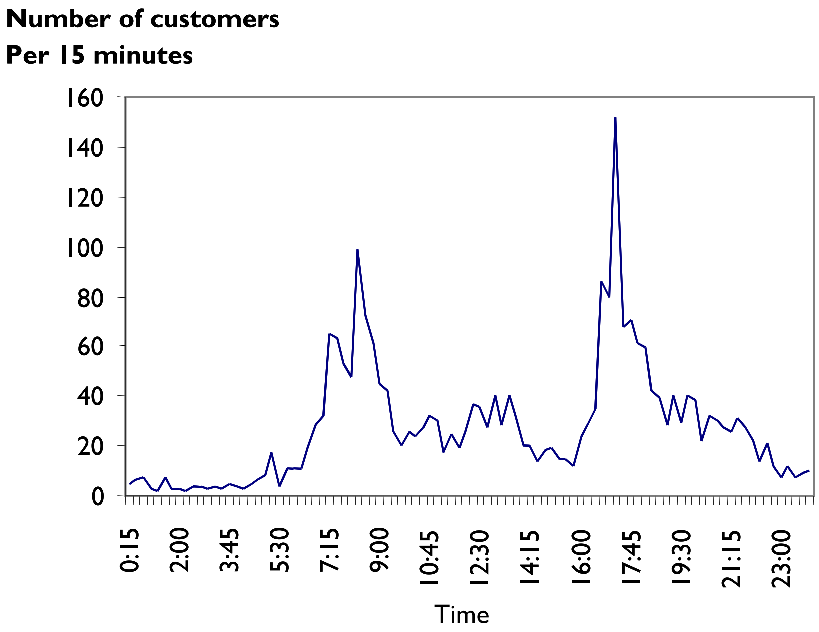
(c) Service and inter-arrival times are random. The *exact* Erlang model assumes that both have a very specific, so-called *exponential distribution;* see next slide. The exponential distribution is fully characterized by a *single parameter*, its mean. In practice, the Erlang model is often used as an approximation , even when the service and inter-arrival time distributions are not exponential

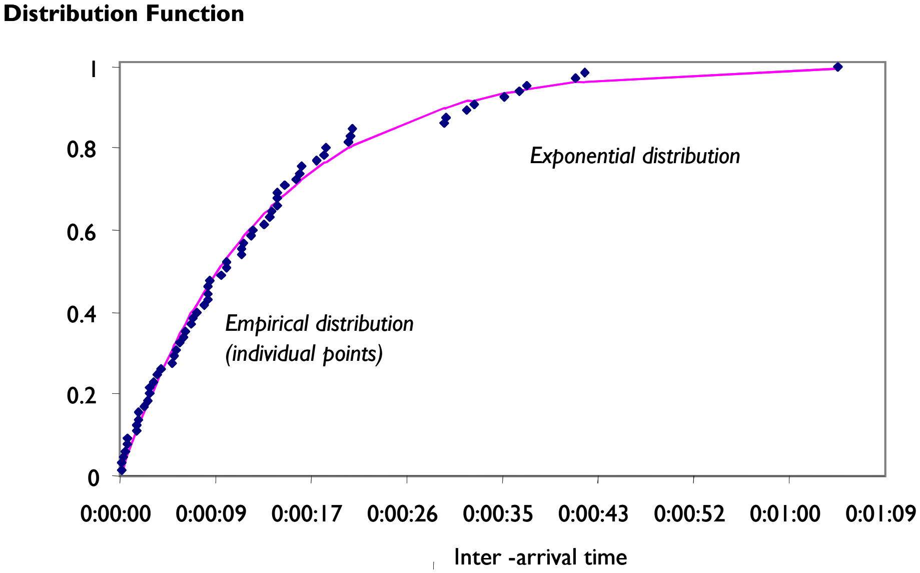
(d) Waiting space is ample

(e) Customers do not leave the system before being served; no abandonments

* Model is sometimes referred to as M/M/s model

Data in Practical Call Center Setting





- Seasonality vs. variability
- Need to slice-up the data

- Within a "slice", time between consecutive calls has exponential distribution.

Offered load: Utilization rate

- Let

s= number of agents/servers

a= λ/μ = offered load= minimum number of agents required

ρ = a/s= utilization rate

Under any kind of randomness, we must have s>a or ρ <1

The difference (s-a) can be thought of as the *service based capacity*. Its magnitude depends on the level of service we want or need to provide:

Erlang Model: Basic Formulas

Service level:

$$\mathrm{SL} = 1 - \mathrm{P_d}\,(\mathrm{s,a})\,\mathrm{e}^{-\,(\mathrm{s-a})\,\mathrm{AWT}\mu} = 1 - \mathrm{P_d}\,(\mathrm{s},\rho)\,\mathrm{e}^{-\mathrm{s}(1-\rho)\,\mathrm{AWT}\,\mu} \qquad (1)$$

Average waiting time:

$$\mathrm{T_q} = \frac{P_d(s,a)}{\mu(s-a)} = \frac{P_d(s,\rho)}{\mu s(1-\rho)} \qquad (2)$$

Probability of delay:

$$\mathrm{P_d}\,(\mathrm{s,a}) = \frac{a^s/s!}{[1-\rho]\left[\sum_{k=0}^{s-1} a^k\!\big/_{k!} + \frac{a^s}{s^{!}}\frac{1}{(1-\rho)}\right]} \qquad (3)$$

Erlang Model: Probability of Delay

Properties of probability of delay P_d:

a) P_d increases from 0 to 1, as ρ increases from 0 to 1

b) For s=1, P_d (1,a)= ρ = a

c) $\text{For } s \geq 10: P_d(s,a) \approx \overline{\Phi}\left(\frac{s - a}{\sqrt{a}}\right)$, with $\Phi(\bullet)$ cdf of standard Normal

d) For given values of a and s, the service level provided does not depend on the absolute waiting standard AWT, but on the relative waiting time standard

$$\text{AWT} * \mu = \text{AWT} / \left(\frac{1}{\mu}\right)$$

= waiting time standard expressed as a percentage of the average service time

Performance Measures when s=1

Probability of delay

P (no delay) $= 1 - \rho$

==> P(delay) = 1 - P(no delay) = ρ

Average time spent in queue

$$T_q = \frac{1}{\mu} \frac{\rho}{1-\rho}$$

Average time spent in <u>system</u>

$$T_s = T_q + \frac{1}{\mu} = \frac{1}{\mu - \lambda}$$

Performance Measures when s=1 (cont'd)

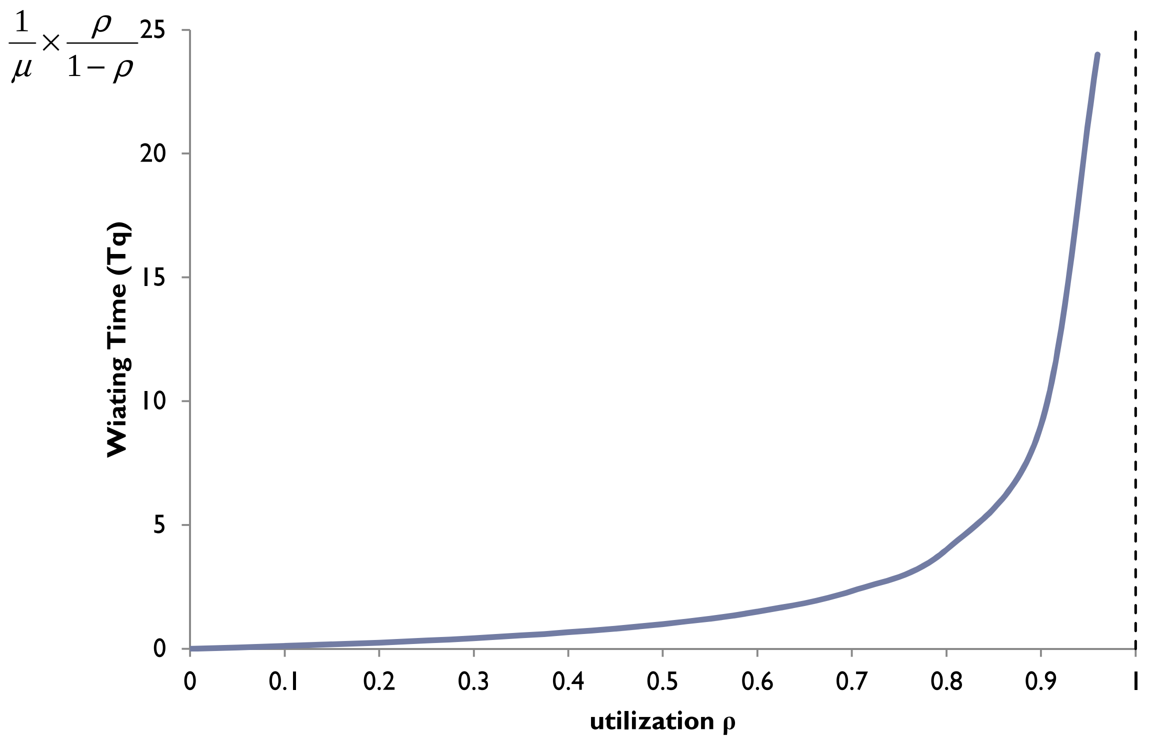
Average queue length

Little's Law => $N_q = \lambda T_q = \dfrac{\rho^2}{1-\rho}$

Average # customers in system

$$N_s = \lambda T_s = \frac{\lambda}{\mu\lambda} = \frac{\rho}{1-\rho}$$

Utilization and Average Waiting Time



Non-Exponential Service and Interarrival Times

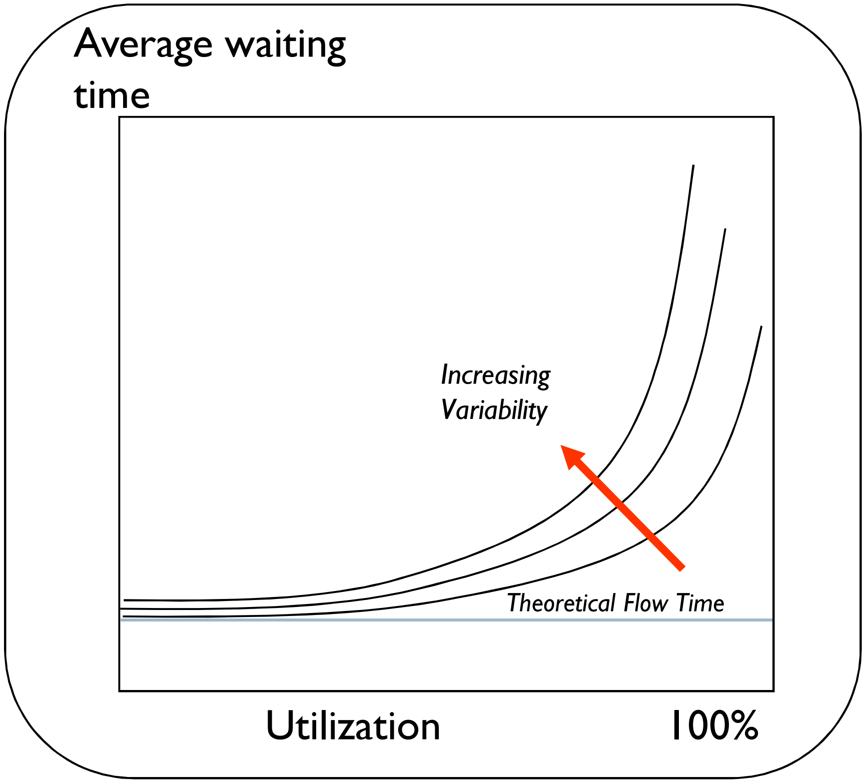
- Variability in service and interarrival times are drivers of waiting time.
- Measured by the *Coefficient of Variation*:

$$CV_s = \frac{\text{Std. Dev. of service time}}{\text{Avg. service time}}$$

$$CV_a = \frac{\text{Std. Dev. of interarrival time}}{\text{Avg. interarrival time}}$$

- Approximation for the avg. waiting time in queue:

$$\underbrace{T_{q,\text{ Erlang}}}_{\text{Waiting time in queue from Erlang}} \cdot \underbrace{\left(\frac{CV_a^{\,2} + CV_s^{\,2}}{2}\right)}_{\text{service time variablility factor}}$$

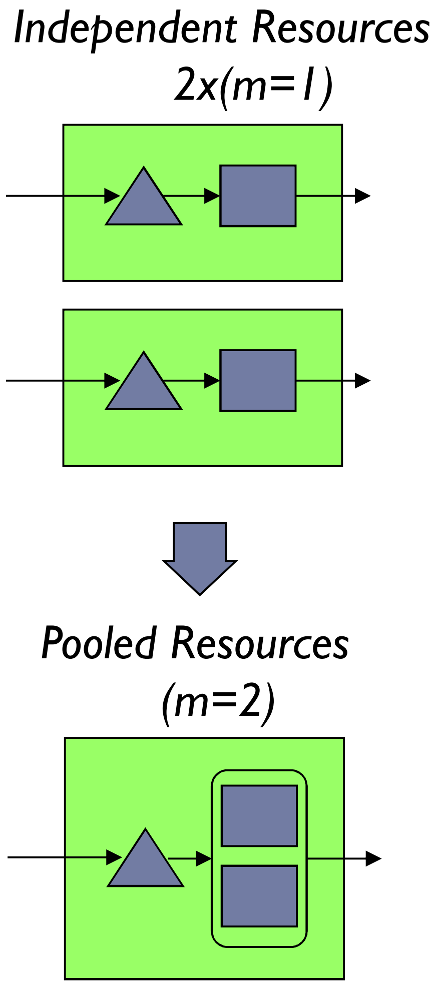


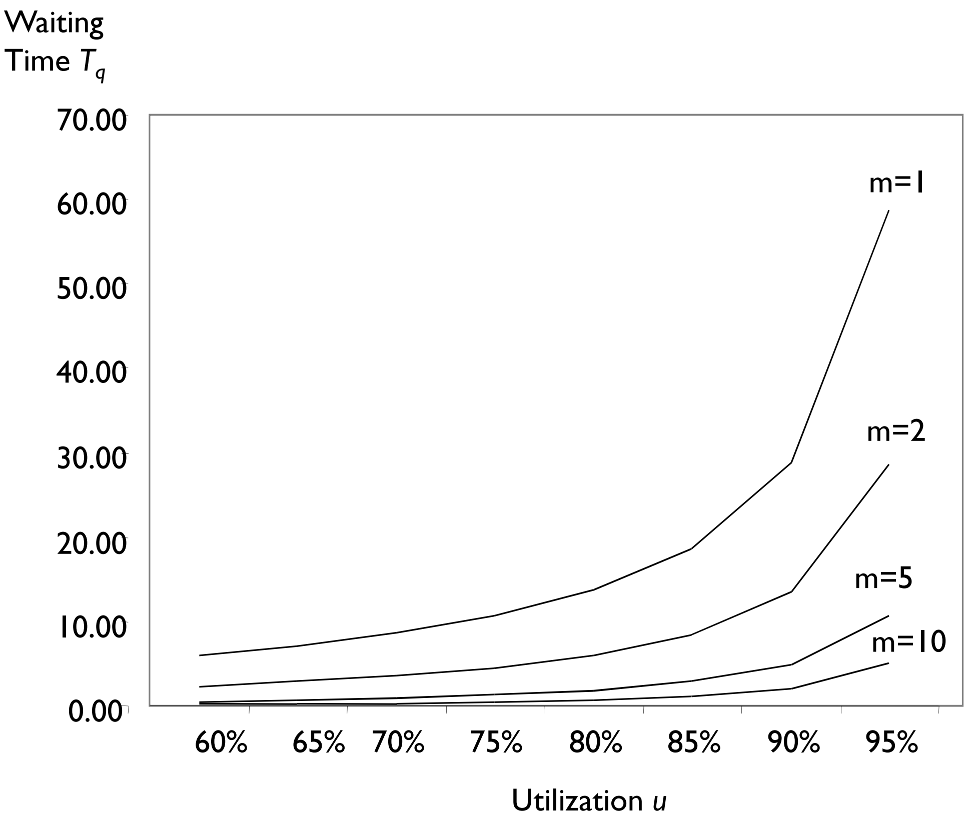
Example: Online retailer

Customers send questions to an online retailer through an on-line chat help desk every 2 minutes, on average, and the standard deviation of the inter-arrival time is also 2 minutes. The online retailer has three employees to answer questions. It takes on average 4 minutes to write a response. The standard deviation of the service times is 2 minutes.

Q: Estimate the average customer wait before being served.

The Power of Pooling





Implications:

(+) Balanced utilization

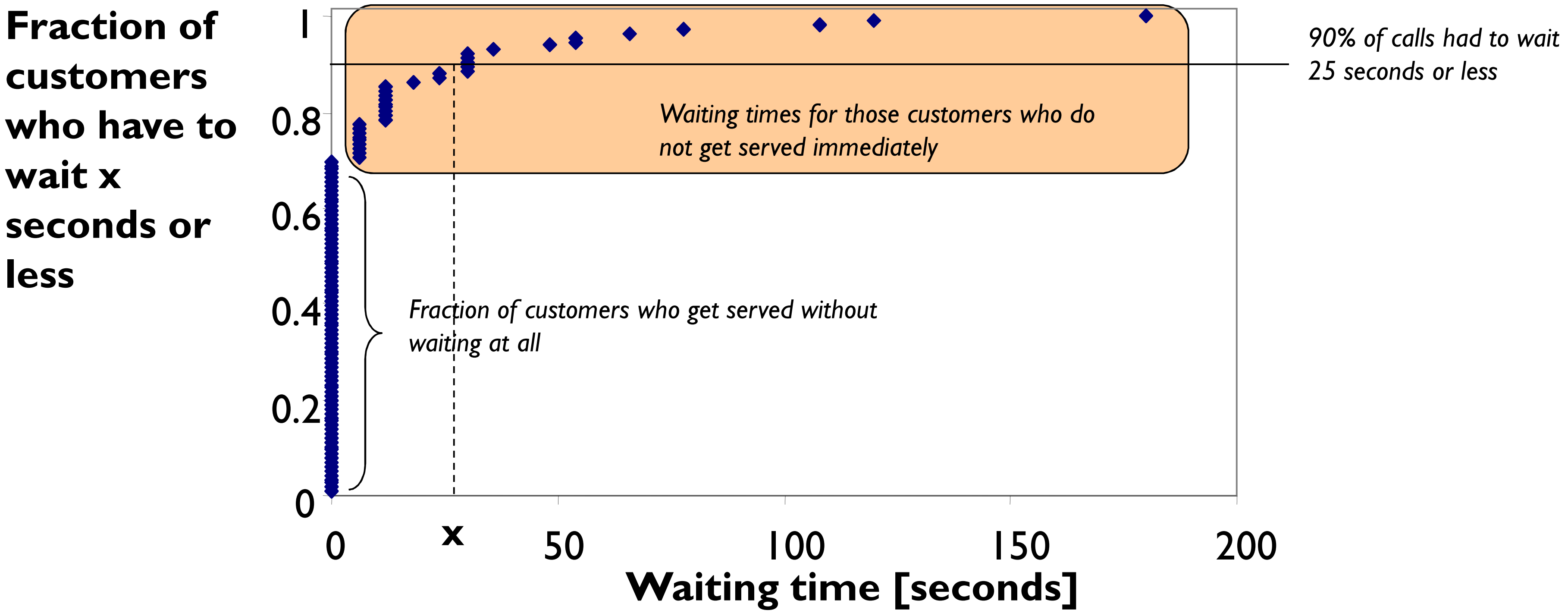
(+) Pool safety capacity

(+) Statistical economies of scale

(-) Change-overs / set-ups

(-) Less specialization

Other measures of performance: Acceptable Wait Time



- **Acceptable Wait Time** (AWT)
- **Service Level** = Probability { Waiting Time ≤ TWT }
- .

Basic Erlang Model: Capacity Analysis

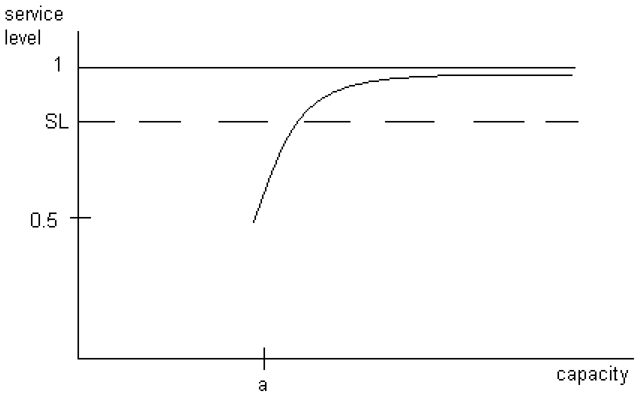
- $$\mathrm{SL} = 1 - \mathrm{P_d}\,(\mathrm{s},\mathrm{a})\,\mathrm{e}^{-(\mathrm{s}-\mathrm{a})\,\mathrm{AWT}\mu} = 1 - \mathrm{P_d}\,(\mathrm{s},\rho)\,\mathrm{e}^{-\mathrm{s}(1-\rho)\,\mathrm{AWT}\,\mu} \qquad (1)$$

- Assume an SLA is given, with given AWT and SL (50%). How much capacity, i.e. how many agents (s) do we need to satisfy the SLA?

- Using the Normal approximation in (3),

$$\mathrm{SL} = 1 - \overline{\Phi}\left(\frac{\mathrm{s}-\mathrm{a}}{\sqrt{\mathrm{a}}}\right)\mathrm{e}^{-(\mathrm{s}-\mathrm{a})\,\mathrm{AWT}\mu}$$

as s increases from a to ∞, SL increases from 0.5 to 1.



Capacity Analysis (cont'd)

- Suppose that we wish to adhere to an agreed SLA (given AWT and SL)
- How does the capacity grow with the demand volume λ?

$$\boxed{s = a + k\sqrt{a}}$$

(Square Root Staffing Formula)

- (k depends on AWT and SL)
- Square Root Staffing Formula shows economies of scale and cost advantages of pooling.

Conclusions

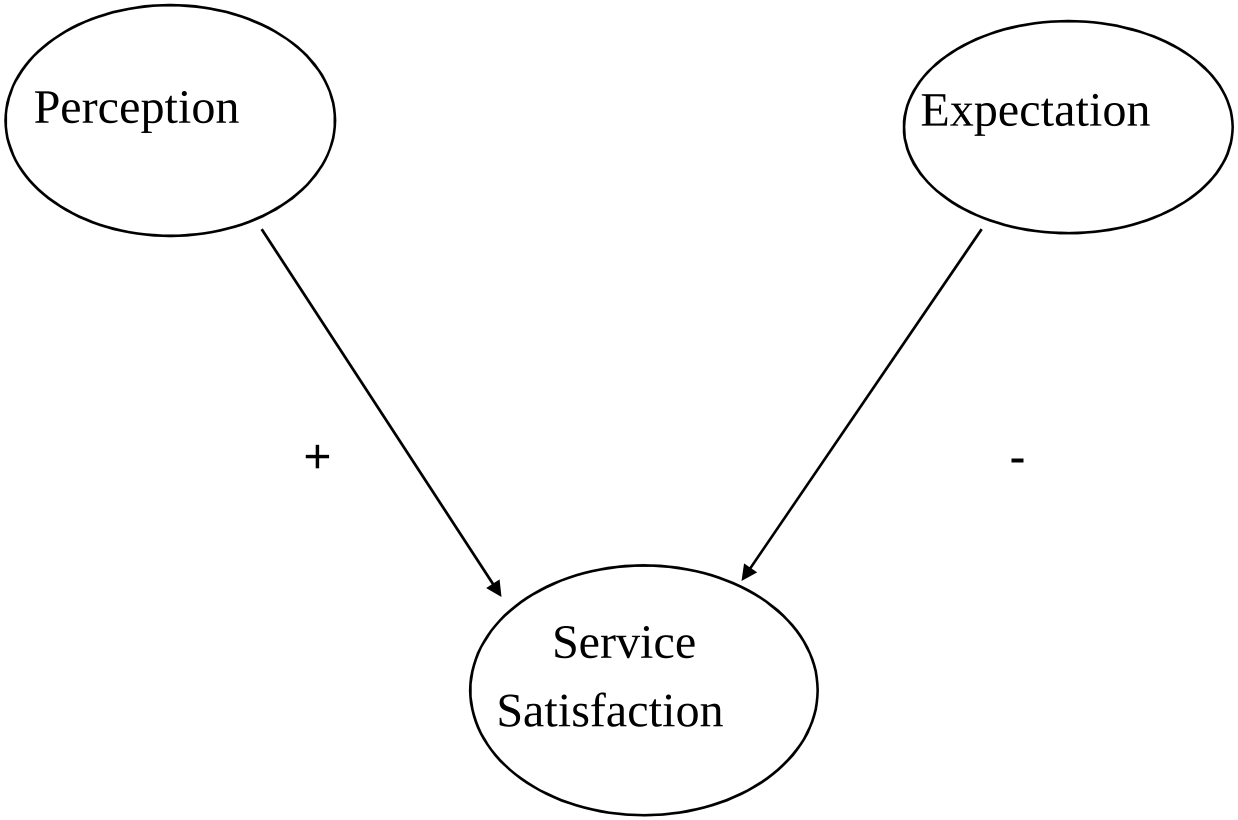
- Variability is the norm, not the exception:
  - Measure, understand the sources and try to reduce it.
  - Accommodate the rest (e.g. by adding capacity).
- Variability leads to waiting times even if utilization<100%.
- Queuing models are useful to:
  - Quantify the effect of variability on system performance.
  - Analyze different scenarios (e.g. reduce average service times, pool servers).

Operations Management: Improving Performance Measures

- Accommodate variability:
  - Increase capacity.
  - Pool servers (statistical economies of scale).
  - Automate or speed up some tasks (e.g., cashier in fast-food restaurants).
  - Pre-process (e.g., fill forms before seeing doctor).
- Manage variability:
  - Reduce variability in arrivals (e.g., appointment systems).
  - Incentives to avoid peak hours (e.g., early-bird special in restaurants).
  - Reduce service time variability (e.g. SOP).
  - Segment customers: express lane in supermarkets.

Perceptions Management: First Law of Service

Satisfaction = Perception - Expectation



Psychology of Queues

1. Unoccupied time feels longer than occupied time
2. Preprocess waits feels longer than in-process waits
3. Anxiety makes waits seem longer
4. Uncertain waits are longer than known, finite waits
5. Unexplained waits are longer than explained waits
6. Unfair waits are longer than equitable waits
7. The more valuable the service, the longer people will wait
8. Solo waiting feels longer than group waiting

Perceptions Management

- Install distractions that entertain and physically involve the customer.
- Keep resources not serving customers out of sight.
- Never underestimate the power of a friendly server.
- Adopt a long-term perspective.